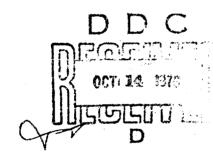# control analysis corporation

D D C

OCT 14 1976

D

headquarters - 800 welch road - palo alto, california - 94304
telephone (415) 326-2100

washington office - 8809 maxwell drive - potomac, maryland - 20854
telephone (301) 299-8488

CONTROL ANALYSIS CORPORATION

800 Welch Road
Palo Alto, California

Technical Report No. 86-22

⑪ August 1976   ⑫ 35p.

⑭ TR-86-22

⑥ VARIANCE REDUCTION FOR REGENERATIVE

SIMULATIONS, I: INTERNAL CONTROL AND

STRATIFIED SAMPLING FOR QUEUES.

⑨ Technical rept.,

by

⑩ Donald L. Iglehart and Peter A. W. Lewis

D D C

⑮ N00014-72-C-0086

⑯ NR-047-106

VARIANCE REDUCTION FOR REGENERATIVE SIMULATIONS, I:

INTERNAL CONTROL AND STRATIFIED SAMPLING FOR QUEUES

by

Donald L. Iglehart
Stanford University
Stanford, California

Peter A. W. Lewis
Naval Postgraduate School
Monterey, California

ABSTRACT

We discuss two methods for reducing the variance of estimators of parameters of limiting distributions of stable stochastic processes in simulations. The methods are discussed in the context of the simple GI/G/1 queue. Of the two methods one, which we call an internal control variable, gives a variance reduction which is roughly uniform over values of the parameters of the process and, in particular, works well for values of ρ_1 the traffic intensity, close to 1.

VARIANCE REDUCTION FOR REGENERATIVE SIMULATIONS, I:
INTERNAL CONTROL AND STRATIFIED SAMPLING FOR QUEUES

by

Donald L. Iglehart   and   Peter A. W. Lewis

1. Introduction and Summary

Simulators are frequently faced with the task of estimating a parameter associated with the limiting distribution of a stochastic process which is being simulated. A methodology based on regenerative processes for obtaining point estimates and confidence intervals for such parameters has recently been developed in Crane and Iglehart (1974a,b), (1975a,b) and Iglehart (1975), (1976a,b). In this paper we shall indicate two techniques, internal control variables and internal stratified sampling, which might be used in conjunction with the regenerative method for obtaining additional variance reduction for the estimates. To illustrate these techniques we shall restrict our attention in this paper to the GI/G/1 queue. Other applications of these ideas will be dealt with in future publications, as well as uses of the regenerative property which may possibly be more suited for obtaining variance reduction for estimates in stable stochastic processes.

In the GI/G/1 queue we assume the zeroth customer arrives at time $t_0 = 0$, finds a free server, and experiences a service time v_0. The $n\underline{th}$ customer arrives at time t_n and experiences a service time v_n. Let the interarrival times $t_n - t_{n-1} = u_n$, $n \geq 1$. Assume the two sequences $\{v_n : n \geq 0\}$ and $\{u_n : n \geq 1\}$ each

consist of independent, identically distributed (i.i.d.) random variables (r.v.'s) and are themselves independent. Let $E\{v_n\} = \mu^{-1}$, $E\{u_n\} = \lambda^{-1}$, and $\rho = \lambda/\mu$, where $0 < \lambda, \mu < \infty$. Thus μ has the interpretation of the mean service rate and λ has the interpretation of the mean interarrival rate. The parameter ρ is called the traffic intensity and is the natural measure of congestion for this system. We shall assume that $\rho < 1$, a necessary and sufficient condition for the system to be stable.

While many characteristics of interest can be estimated using the regenerative method, we shall restrict our attention to the waiting time of the $n\underline{\text{th}}$ customer, W_n (time from arrival to commencement of service). To obtain a representation for the process $\{W_n : n \geq 0\}$ let $X_n = v_{n-1} - u_n$ and set $S_0 = 0$, $S_n = X_1 + \ldots + X_n$, $n \geq 1$. The following well-known recursive relationship exists for the W_n's:

$$W_0 = 0, \quad W_{n+1} = [W_n + X_{n+1}]^+, \quad n \geq 0.$$

By induction, we also have

$$W_n = \max\{S_n - S_k : k = 0, 1, \ldots, n\}, \quad n \geq 0.$$

Using the strong Markov property of the process $\{S_n : n \geq 0\}$ it can be shown that there exists a sequence of integer-valued r.v.'s $\{\beta_k : k \geq 0\}$ such that $\beta_0 = 0$, $\beta_k < \beta_{k+1}$, and $W_{\beta_k} = 0$ with probability one. In other words, the customers numbered β_k are those lucky fellows who arrive to find a free server and experience no waiting in the queue. The fact that there exists an infinite number of such customers in the GI/G/1 queue is a direct consequence of the

assumption that $\rho < 1$ and the strong law of large numbers. If we let $\alpha_k = \beta_k - \beta_{k-1}$, $k \geq 1$, then α_k represents the number of customers served in the $k^{\underline{th}}$ busy period (b.p.) and they are numbered $\{\beta_{k-1}, \beta_{k-1}+1, \ldots, \beta_k-1\}$.

Next define the sequence $\{Y_k: k \geq 1\}$ by

$$Y_k = \sum_{j=\beta_{k-1}}^{\beta_k-1} W_j,$$

the sum of the waiting times in the $k^{\underline{th}}$ b.p. Since the queue is stable for $\rho < 1$ we have $W_n \Rightarrow W$ as $n \to \infty$, where $\Rightarrow$ denotes weak convergence.

<u>Our goal is to estimate $E\{W\}$ by simulation.</u>

It is known that $E\{W\} < \infty$ if and only if $E\{(X_1^+)^2\} < \infty$. We shall, for simplicity, assume that $E\{v_0^2\} < \infty$ which will guarantee that $E\{W\} < \infty$. The regenerative simulation method is based on the analytic results that the sequence $\{(Y_k, \alpha_k): k \geq 1\}$ is independent and identically distributed and $E\{W\} = E\{Y_1\}/E\{\alpha_1\}$, the ratio of two means. This suggests using the ratio of estimates of $E(Y_1)$ and $E(\alpha_1)$ to estimate $E(W)$. Thus, if we let $\bar{Y}(n) = \frac{1}{n} \sum_{k=1}^{n} Y_k$ and $\bar{\alpha}(n) = \frac{1}{n} \sum_{k=1}^{n} \alpha_k$, where n now denotes the number of cycles observed, then a ratio estimate of $E(W)$, for example, is

$$W(n) = \bar{Y}(n)/\bar{\alpha}(n). \qquad (1.1)$$

(In the sequel we drop the dependence on n unless necessary.)

Now let $Z_k = Y_k - E\{W\}\alpha_k$, $k \geq 1$, and note that the sequence $\{Z_k: k \geq 1\}$ is i.i.d. and $E\{Z_k\} = 0$. We assume the variance of Z_k,

3

$$E\{Z_k^2\} = \sigma^2 = \text{var}\{Y_k\} - 2 \text{ cov}\{Y_k,\alpha_k\}E\{W\} + \text{var}\{\alpha_k\}E^2\{W\} \qquad (1.2)$$

is finite and positive. Then one can easily show that as $n \to \infty$

$$\frac{n^{1/2}[\bar{Y}(n)/\bar{\alpha}(n) - E\{W\}]}{\sigma/E\{\alpha_1\}} \Rightarrow N(0,1), \qquad (1.3)$$

where $N(0,1)$ is a mean zero, variance one normal random variable.
This result yields a confidence interval for $E\{W\}$ provided we
can estimate $\sigma/E\{\alpha_1\}$. A variety of estimates have been studied
and are reported on in Iglehart (1975a). Here we just mention two.
The so-called classical estimate for $\sigma/E\{\alpha_1\}$ is given by

$$\hat{S}_1 = [S_{11} - 2S_{12}(\bar{Y}/\bar{\alpha}) + S_{22}(\bar{Y}/\bar{\alpha})^2]^{1/2}/\bar{\alpha},$$

where S_{11} is the sample variance of the Y_k's, S_{22} of the α_k's,
and S_{12} the sample covariance of the Y_k's and α_k's.

The second estimate of $\sigma/E\{\alpha_1\}$ is the jackknife estimate
which is defined to be

$$\hat{S}_2 = [\sum_{i=1}^{n} (\theta_i - \bar{\theta})^2/(n-1)]^{1/2},$$

where $\theta_i = n(\bar{Y}/\bar{\alpha}) - (n-1)(\sum_{j \neq i} Y_j / \sum_{j \neq i} \alpha_i)$, $1 \leq i \leq n$, and $\bar{\theta} = \frac{1}{n}\sum_{i=1}^{n} \theta_i$
is called the jackknifed estimator of $E(W)$. Both $\hat{S}_1$ and $\hat{S}_2$
are strongly consistent.

The jackknifing technique is known to work particularly
well as a confidence interval estimate for ratios; for a large
number of cycles n the computational problem is severe, but in
that case the technique using (1.3) and $\hat{S}_1$ works well. For
details see Miller (1974) and Iglehart (1975).

4

The problem addressed in this paper is how to apply variance reduction techniques with the ratio estimator $\bar{W}(n)$. In almost all practical situations, where in particular one might want to compare the mean waiting times of two different queueing systems (Iglehart (1976)), there is a premium on achieving the minimum possible variance for the estimation in the given computing time (number of cycles allowed). Variance reduction techniques for simulations are discussed in Kleijnen (1974) and Gaver and Thompson (1973), but there are difficulties in applying these to ratio estimates and regenerative systems. In particular variance reduction via the usual control variable techniques is difficult. A variation of this technique, which we call internal control variables and which is generally useful for ratio estimates, is introduced and shown to give considerable variance reduction for the point estimate $\bar{W}(n)$. Another technique, internal stratified sampling, is also explored. It is a natural technique to use but appears to be difficult to use with ratios. Moreover, it becomes less and less effective as $\rho \to 1$, while the internal control technique holds up well for ρ close to 1. In fact the internal control variables described here for an M/M/1 queue give a variance reduction which is fairly uniform for all values of μ and ratios $\rho < 1$. Better results can be obtained for particular cases of the parameter ρ.

It will also be apparent after the development of the variance reduction techniques in the next sections that the two techniques for confidence interval estimates discussed above apply to the estimates after variance reduction has been applied.

5

2. Internal Control

Two main methods for variance reduction, antithetic variates and control variables (Gaver and Thompson (1973)) have been put forward for use with queues in the non-regenerative situation. Of these, antithetics has very limited utility beyond the simple M/M/1 situation in which it is patently clear how to generate two realizations of samples which give negatively correlated estimates. That scheme for generating negatively correlated estimates does not work with the regenerative method because the regenerative blocks in the original realization of the queue and the antithetic realization get out of synchronization. No alternative way has been found to utilize antithetic variates with the regenerative technique and we feel, along with many computer scientists, that the technique has limited validity in systems simulation.

The technique of using a control variable and, in particular, a regression-adjusted control variable (Gaver and Thompson (1973) p. 587) is much more broadly applicable in systems simulation, although it is again difficult to adapt to the regenerative situation. Briefly, say we are simulating an M/G/1 queue to estimate E(W) with the non-regenerative technique of averaging the first m waiting times to obtain an estimate of the unknown E(W),

$$\widetilde{W}_m = \frac{1}{m} \sum_{j=1}^{m} W_j . \tag{2.1}$$

The same random numbers used to generate the m non-exponentially distributed service times are used to generate m exponential service times for simulating an M/M/1 queue whose input stream

6

is identical to that of the simulated M/G/1 queue. One would
then expect that if the G-distribution is not too different from
the exponential distribution, successive waiting times, say W_j',
in the M/M/1 queue realization would be close to (positively
correlated with) the corresponding waiting times W_j in the
M/G/1 queue. Consequently, the average of the W_j''s, say $\widetilde{W}_m'$,
will be positively correlated with $\widetilde{W}_m$ and one can form a new
estimate

$$\widetilde{\widetilde{W}}_m = \widetilde{W}_m + \beta(\widetilde{W}_m' - E(\widetilde{W}_m')). \qquad (2.2)$$

Now $E(\widetilde{W}_m')$ is close to $E\{W'\} = \rho/\mu(1-\rho)$ for m large, so
$E(\widetilde{\widetilde{W}}_m) \approx E(\widetilde{W}_m)$ and the variance of the new estimate will be a minimum
if

$$\beta = -cov(\widetilde{W}_m, \widetilde{W}_m')/var(\widetilde{W}_m'); \qquad (2.3)$$

in fact

$$\frac{\sigma_{\widetilde{\widetilde{W}}_m}}{\sigma_{\widetilde{W}_m}} = \frac{s.d.(\widetilde{\widetilde{W}}_m)}{s.d.(\widetilde{W}_m)} = (1-r^2)^{1/2}, \qquad (2.4)$$

where

$$r = corr(\widetilde{W}_m, \widetilde{\widetilde{W}}_m) = \frac{cov(\widetilde{W}_m, \widetilde{\widetilde{W}}_m)}{\sigma_{\widetilde{\widetilde{W}}_m} \sigma_{\widetilde{W}_m}}. \qquad (2.5)$$

There are a number of important points to be noted about
this technique:

(i)   It allows one to use known analytical results (such as
      the expected value of the limiting waiting time in an

7

M/M/1 queue) to reduce the variance of a simulation. Such use of analytical results is a basic principle of simulation.

(ii) It can be extended to non-linear controls or to multiple control by more than one control variable.

(iii) The great art in the technique lies in finding a control whose mean value is known and which is highly correlated with the estimator which is being controlled. Thus (2.4) says that $|r|$ must be close to 0.9 to reduce the standard deviation $\sigma_{\widetilde{W}_m}$ to one half of $\sigma_{\widetilde{W}_m}$ and this generally is equivalent to reducing the required sample size for a given precision by a quarter. Controls which are that highly correlated with the estimate are not easy to find.

(iv) It is in general too much to ask that the correlation and variances in (2.3), (2.4), and (2.5) be known analytically. They, therefore, must be estimated from the simulation data and this will reduce the variance reduction which is attained.

Now using this method to control the regenerative estimate given in the introduction,

$$W(n) = \bar{Y}(n)/\bar{\alpha}(n),$$

where n refers to a fixed number of cycles in the queue, one runs into the same problem of synchronization as with the antithetic case, namely n cycles in the M/G/1 queue may take a very different

8

number of waiting times to achieve than are needed for n cycles in the M/M/1 queue. Moreover, the correlation between Y_k' and Y_k will be weak and made even weaker by the use of the ratio estimate. This problem becomes rapidly apparent in simulation studies of the technique.

The following technique which we have called internal (within block) control has been developed to overcome this. It is, in fact, a special case of the very broad technique called concomitant variables in Gaver and Thompson (1973), p. 588 and will be illustrated only for point estimation of $E(W)$ and $E(\alpha)$. Its extension to other quantities discussed by Crane and Iglehart (1974a) is immediate in principle, although the control quantities discussed below may be different.

2.1 Internal Control: Basic Ideas

The idea of an internal control variable is simple. In the estimate $W(n)$, the averages $\bar{Y}(n)$ and $\bar{\alpha}(n)$ contain n random variables Y_k and α_k which are each functions only of the α_k interarrival and service times occurring in the $k\underline{\text{th}}$ cycle (or b.p.) and are independent of the other interarrival times and service times. Thus, it is natural to use some function of these $2\alpha_k$ random variables to control each Y_k and α_k.

The naive application of this idea is that if we can reduce the variance of both the numerator and denominator $\bar{Y}(n)$ and $\bar{\alpha}(n)$ we will reduce the variance of $W(n)$, but we will show that the situation is more complex than this. We will denote a function of the random variables in the $k\underline{\text{th}}$ cycle by $C(k)$ but will generally

9

drop the index. In general we also use $C_T(k)$ to denote a control for the numerator (t⸱) in the ratio estimator and $C_B(k)$ to denote a control for the denominator (bottom).

Typically, $C(k)$, or simply C, could be the difference between the service time v_{β_k} and the time to arrival of the next customer from the arrival of the $\beta_k\underline{th}$ customer, namely $\mu_{\beta_{k+1}}$. This difference has, in a GI/G/1 queue, a known mean $\mu^{-1} - \lambda^{-1}$ and large positive values of this function $C(k)$ correspond to large values of Y_k and α_k, and vice versa. We return to specific control variables and their computation in the next sub-section.

Note that one can control either the top or bottom of the ratio estimator, or both; to fix ideas assume we control the top and have, in general, an internally controlled estimate

$$W_{CT}(n) = \frac{\frac{1}{n}\sum_{k=1}^{n}\{Y_k + \beta_T(C_T - E(C_T))\}}{\bar{\alpha}}, \qquad (2.6)$$

where β_T, as in the usual control estimation technique, is fixed so as to minimize $\mathrm{var}(W_{CT}(n))$. In practice it is usually estimated from the simulation data.

Now the quantity $\sigma^2/E^2\{\alpha\}$, where $\sigma^2 = E\{Z_k^2\}$ is given at (1.2), is just the leading term in the asymptotic expansion for the variance of the ratio estimator $W(n)$, and carries over to the more complicated situation (2.6) to give (asymptotically as $n \to \infty$)

$$n\,\mathrm{var}(W_{CT}(n)) \to \sigma_T^2 = \left\{\frac{E(Y)}{E(\alpha)}\right\}^2 \{\mathrm{var}(Y' + \beta_T C_T' - \alpha')$$

$$= \left\{\frac{E(Y)}{E(\alpha)}\right\}^2 \mathrm{var}\{(Y' - \alpha') + \beta_T C'_T\}, \qquad (2.7)$$

10

where $Y' = Y/E\{Y\}$, $\alpha' = \alpha/E\{\alpha\}$, and $C_T' = C_T/E\{Y\}$. For a derivation of this result see Cramér (1946), p. 354, eq. (27.7.3).

It follows from (2.7), as before, that to minimize the asymptotic variance of $W_{CT}(n)$ we must take

$$-\beta_T = \frac{\text{cov}(Y'-\alpha',C_T')}{\text{var}(C_T')} = \frac{\text{cov}(Y,C_T)}{\text{var}(C_T)} - \frac{E(Y)}{E(\alpha)}\frac{\text{cov}(\alpha,C_T)}{\text{var}(C_T)} . \qquad (2.8)$$

But most importantly we notice that C_T' must be highly correlated with the difference $Y' - \alpha'$. To achieve this is much more difficult than finding a quantity which is highly correlated with either of Y or α, simply because Y and α are highly correlated and increase together. In particular if $\alpha = 1$, which has a high probability if ρ, the traffic intensity is small, then $Y = 0$.

Note, too, that $E(W) = E(Y)/E(\alpha)$, the quantity we are trying to estimate in the simulation, appears in (2.8) for the top control regression coefficient. Also similar equations to (2.7) and (2.8) pertain to the case where the bottom of the ratio is controlled (in this case $\bar{\alpha}$), and simultaneous equations can be derived for β_T and β_B if both top and bottom are controlled. The additional complexity in estimating β_T and β_B does not seem to be justified by simulation results (discussed later) which also show that if only one control is used there seems little to choose in terms of reduction achieved between putting it on the top of the bottom. In both cases the control must be highly correlated with the difference between Y and α.

The above results are generally applicable for any regenerative process in which ratio estimates are used. The choice of C is, of course, the art in the design of a simulation with variance reduction and is considered for the GI/G/1 and specifically the M/M/1 queue in the next section. In all cases this choice is limited by one's ability to compute analytically E(C).

2.2 Internal Control: Design Considerations

We discuss here the design of internal controls for the M/M/1 queue, the ideas being applicable to the GI/G/1 case with the proviso that the computations might be considerably more difficult. Simple computations of E(C) are given here and more difficult ones in the Appendix; we do not distinguish between bottom and top controls, since both must be correlated with the difference $Y - \alpha$, and we drop consideration of cycle number, since all variables are within the cycles which have identical structure.

Again, we are considering estimation of E(W), but of the many possible controls, those listed below would probably work as well with other functions of W, e.g. percentiles. The controls are listed roughly in order of complexity of computation and of supposed correlation with $Y - \alpha$. This can usually only be guessed at and generally the more elaborate controls which might have greater correlation with $Y - \alpha$ are more difficult computationally.

Superscripts on C are labels to differentiate the controls. We have discussed above the difference $X_1 = v_0 - u_1$, whose moments are simple to compute. Then we have

12

$$C^{(1)} = X_1^+ = W_1 = 0 \quad \text{if} \quad \alpha = 1 \qquad (2.9)$$

$$= X_1 \quad \text{if} \quad \alpha \geq 2. \qquad (2.10)$$

It is easy to compute $E(C^{(1)})$ for $M/G/1$ queues and possible for the $GI/G/1$ queue. Thus, for the $M/M/1$ case we have, using the Markov property of the exponential distribution,

$$P\{v_0 < u_1\} = P\{\alpha = 1\} = \int_0^\infty [1 - F_u(y)] dF_v(y)$$

$$= \int_0^\infty e^{-\lambda y} (\mu e^{-\mu y}) dy = \frac{1}{1+\rho} \qquad (2.11)$$

which goes from 1 to 1/2 as $\rho = \lambda/\mu$ goes from 0 to 1; furthermore,

$$P\{\alpha \geq 2\} = 1 - \frac{1}{1+\rho} = \frac{\rho}{1+\rho} \ . \qquad (2.12)$$

Now given that $X_1 = v_0 - u_1$ is greater than zero, the excess $v_0 - u_1$ is distributed as an exponential random variable with parameter μ. Therefore

$$E(X_1^+) = E(C^{(1)}) = 0 \times \frac{1}{1+\rho} + \frac{\rho}{1+\rho} \frac{1}{\mu} = \frac{1}{\mu}\left(\frac{\lambda}{\lambda+\mu}\right) \ . \qquad (2.13)$$

The variance of $C^{(1)}$ can also be computed.

One would generally like to obtain more correlation of C and $Y - \alpha$ when Y is large, and one feels this can be done by bringing in the second waiting time. Thus we have as control candidates

13

$$c^{(2)} \quad \begin{aligned} &= 0, &&\text{if } \alpha = 1 \\ &= X_1^+ + X_2^+, &&\text{if } \alpha \geq 2; \end{aligned} \tag{2.14}$$

$$c^{(3)} = (X_1^+ + X_2)^+ = W_2; \tag{2.15}$$

$$c^{(4)} = \begin{aligned} &0, &&\text{if } \alpha = 1 \\ &X_1^+ + (X_1^+ + X_2)^+, &&\text{if } \alpha \geq 2. \end{aligned} \tag{2.16}$$

The control $c^{(2)}$ is $W_0 = 0$ if $\alpha = 1$, it is $W_0 + W_1$ if $\alpha = 2$, and it is $W_0 + W_1 + X_2^+$ if $\alpha \geq 3$. It is an attempt to capture the effect of the first two waiting times without the additional computational complexities involved in computing the expectations of $c^{(3)}$ and $c^{(4)}$.

Simulation results show that $c^{(3)}$ and $c^{(4)}$ give very little more control than $c^{(2)}$ for which, in the M/M/1 case, we have

$$E(c^{(2)}) = 0 + E(X_1^+ + X_2^+, \alpha \geq 2)$$

$$= \frac{\rho}{1+\rho} \{ E(X_1^+ + X_2^+ \mid X_1^+ > 0 \}$$

$$= \frac{\rho}{1+\rho} \left\{ \frac{1}{\mu} + E(X_2^+ \mid X_1^+ > 0) \right\}$$

$$= \frac{\rho}{1+\rho} \left\{ \frac{1}{\mu} + E(X_2^+) \right\} = \frac{\rho}{1+\rho} \left\{ \frac{1}{\mu} + \frac{\rho}{1+\rho} \frac{1}{\mu} \right\} \tag{2.17}$$

using (2.12) and the fact that X_2^+ is independent of X_1^+. We use the notation $E\{X, A\}$ for $E\{X 1_A\}$, where X is a random variable, A an event, and 1_A the indicator function of A.

Similar computations go through for $c^{(3)}$ and $c^{(4)}$; from these we will need later the following illustrative result (M/M/1 case).

14

$$P\{\alpha=2\} = P\{X_1^+>0, X_1^++v_1 \leq u_2\}$$

$$= P\{X_1^+>0, v_0-u_1+v_1 \leq u_2\}$$

$$= \frac{\rho}{1+\rho}\{P(u_2 \geq Y\} = \frac{\rho}{1+\rho} \cdot \frac{\mu^2}{(\mu+\lambda)^2} = \frac{\rho}{(1+\rho)^3} \ , \qquad (2.18)$$

where Y, the sum of two exponential(μ) random variables, has
a Gamma$(\mu,2)$ distribution.

From this we also get that

$$P\{\alpha \geq 3\} = 1 - \frac{1}{1+\rho} - \frac{\rho}{(1+\rho)^3} = \frac{\rho^2(2+\rho)}{(1+\rho)^3} \ . \qquad (2.19)$$

Now, none of these four controls is specifically designed
to be correlated with the difference $Y-\alpha$. As a result they work
well for μ and λ such that Y takes on values much larger
than α.

To fix this one might try

$$c^{(5)} = c^{(i)} - \alpha, \qquad i = 1,2,3 \text{ or } 4, \qquad (2.20)$$

since the mean $E(c^{(5)})$ is easily calculated if $E(c^{(i)})$ is
known for $i = 1,2,3,$ or 4, and $E(\alpha)$ is known. But $E(\alpha)$ is
$1/(1-\rho)$ for any $M/G/1$ queue (Cohen, 1969); approximations for
the $GI/M/1$ case are discussed in the Appendix.

There is an additional problem of dimensional stability
involved in using $c^{(5)}$, as with $c^{(1)}, c^{(2)}, c^{(3)}, c^{(4)}$, in that
$E(c^{(i)})$ depends on both μ and ρ, while $E(\alpha)$ depends only on
ρ. Thus control is not uniform across the whole range of param-
eter values.

15

To avoid this one use

$$c^{(6)} = c^{(i)} - \alpha/\mu. \qquad (2.21)$$

This, with $i = 2$, was found, in simulation studies, to be the most successful control variable in that it obtained a variance reduction which was uniform in μ and ρ (or μ and λ) and its mean value is fairly simple to compute. These simulation results are discussed in the next subsection.

Note that multiple control variables using any of the above controls can be used. In particular, one need not take the difference of say, $c^{(2)}$ and α/μ, but may use a multiple control. However, the fact that two regression coefficients, say $\beta_T^{(1)}$ and $\beta_T^{(2)}$ must be estimated from the simulation data makes the possible gain in variance reduction of dubious value.

Note also that since all the control comes from within cycles, there is no reason that the confidence interval estimation iques referred to in the introduction would not go through for the variance-reduced estimates. This has been verified in simulation runs.

2.3 Internal Control: Simulation Results

It is not possible to verify analytically what variance reduction will be obtained via the several internal controls listed in the previous section, or to get an idea of the magnitude of the effect. Even for something as simple as $c^{(1)}$ it is difficult to compute analytically the correlation between $c^{(1)}$ and $Y - \alpha$ for the $M/M/1$ queue, and this is what is required in the equation (2.5) to find the variance reduction.

Thus, we resorted to simulations to verify the amount of variance reduction obtained and the relative effectiveness of the various controls. In the final simulations all runs were performed on IBM System 360/67 computer using the LLRANDOM package (Learmonth and Lewis (1973)) which generates random numbers according to the scheme given by Lewis, Goodman, and Miller (1969) and exponentially distributed random numbers using the Marsaglia "rectangle-wedge-tail" method. Tests of the random number generator are given in Learmonth and Lewis (1974).

Of the extensive simulation checks performed, we give here only a summary of the conclusions and one detailed tabulation and one short tabulation in the case of the most suitable control.

(1) The controls $c^{(2)}$, $c^{(3)}$ and $c^{(4)}$ do much better generally than $c^{(1)}$, with little improvement over $c^{(2)}$ obtained by use of $c^{(3)}$ and $c^{(4)}$. We say generally because results vary with λ and μ and their ratio ρ.

(2) Subtracting the number of customers served in a busy period generally improves the variance reduction. By making it dimensionally stable as in (2.21) with $i = 2$ we obtain a "variance reduction" measured in terms of ratios of standard deviations, of approximately 70%, uniformly over λ and μ. This is roughly equivalent to halving the number of cycles (b.p.'s) that one must simulate; $(0.7)^2 \approx .5$. Much better reductions can be obtained for smaller ρ (i.e. $\rho = 0.25$) by specially designed controls; the point is that $c^{(6)}$ using $c^{(2)}$ works even out at $\rho = 0.99$, where variance reduction is extremely important.

17

Table 1 shows results obtained by simulating an M/M/1 queue out to n = 2000 cycles and replicating the simulation 250 times to estimate the variance of the estimates $W(n)$, $W_{CT}(n)$, $W_{CB}(n)$, where we drop the n for convenience. Here, we have specifically that

$$W_{CB}(n) = \frac{\frac{1}{n} \sum_{k=1}^{n} Y_k}{\frac{1}{n} \sum_{k=1}^{n} [\alpha_k + \beta_B (C^{(6)} - E\{C^{(6)}\})]} \, ,$$

where

$$C^{(6)} = C^{(2)} - \alpha/\mu. \tag{2.22}$$

The estimated precision (standard deviations) of the estimates of E(W) are given in brackets under the estimates.

The results in Table 1 are for $\rho = 0.5$ and three values of μ; the results are not very different at different values of ρ. The case $\rho = 0.99$ is given in Table 2. The second, third and fourth columns in the Tables give correlations between the control and $Y - \alpha$ etc., from which the theoretical variance reduction can be computed. They are very close to the values given in the next to last column, from which we deduce that estimating β_T and β_B affects the variance reduction only slightly. There is negligible effect of different values of μ for fixed $\rho = 0.5$ and fixed $\rho = 0.99$.

Note that for the results for $\rho = 0.99$ given in Table 2, the variance reduction is 73% (about the same as for $\rho = 0.5$). For the

18

case where the control is on the top, i.e. for $\bar{Y}$, the variance reduction is not quite as good for control of α. Note too that the estimated values appear in some cases to be at least three or four standard deviations from the true mean. This is because the estimates $\bar{W}$, $\bar{W}_T$ and $\bar{W}_B$ can be seen from the 100 replications to be non-normal. In other words, for high $\rho(0.99)$, the simulation needs to be taken out further than 2000 cycles.

Table 1. Internally controlled ratio estimates of E(W) in an M/M/1 queue. Change in correlations and variance reduction due to change of scale for fixed $\rho = 0.5$. Number of cycles, 2,000; number of replications of the simulation, 250. The percentage column gives the percentage reduction in standard deviation of the estimate of E(W) due to control, over the standard deviation of the uncontrolled estimate. Control C is $C^{(6)}$ given in (2.22).

μ	$r_{C;Y'-\alpha'}$	$r_{Y,\alpha}$	$r_{Y,C}$	Av W, Av(W_{CT}), Av(W_{CB})	var(W), var(W_{CT}), var(W_{CB})	%	True E(W)
0.0500	0.73390	0.86665	-0.81991	10.00246 (0.05599)	0.78392	100%	10.000
				9.95958 (0.03832)	0.36713	72%	10.0000
				9.99513 (0.03876)	0.37553	71%	10.0000
0.50	0.73392	0.86666	-0.81993	1.00026 (.00560)	0.00784	100%	1.00000
				0.99600 (.00384)	0.00369	72%	1.00000
				0.99955 (0.00388)	0.003370	71%	1.00000
5.000	0.73392	0.8666	-0.81993	0.10003 (0.000560)	0.0000784	100%	0.00000
				0.09960 (0.000560)	0.0000369	72%	0.10000
				0.09995 (0.0003833)	0.0000377	71%	0.10000

Table 2. Internally controlled ratio estimates of $E(W)$ in an M/M/1 queue. Change in correlations and variance reduction due to change of scale for fixed $\rho = 0.99$. Number of cycles, 2,000; number of replications of the simulation, 100. The percentage column gives the percentage reduction in standard deviation of the estimate of $E(W)$ due to control, over the standard deviation of the uncontrolled estimate. Control C is given in (2.22).

μ	$r_{C;Y'-\alpha'}$	$r_{Y,\alpha}$	$r_{Y,C}$	Av $\bar{w}$ Av($\bar{w}_{CT}$) Av($\bar{w}_{CB}$)	var($\bar{w}$) var($\bar{w}_{CT}$) var($\bar{w}_{CB}$)	%	True $E(W)$
0.10	0.78524	0.95775	-0.95780	8.80116 (0.27489)	7.5566	100%	10.00
				9.20920 (0.20023)	4.0090	73%	10.00
				9.71925 (0.22302)	4.9739	81%	10.00
1.00	0.78521	0.95774	-0.95779	98.04502 (2.7521)	757.393	100%	100.00
				92.11951 (2.0150)	406.005	73%	100.00
				97.13507 (2.2365)	500.218	81%	100.00
10.00	0.78516	0.95773	-0.95779	880.255 (27.496)	75604.48	100%	1000.00
				921.032 (20.080)	40320.21	73%	1000.00
				971.474 (22.323)	49831.35	81%	1000.00

3. Internal Stratified Sampling

Another technique for variance reduction which can be potentially used with the regenerative method is stratified sampling. For a brief description see Kleijnen (1974), p. 110. In essence this uses analytical information in the following way.

If we can stratify or partition a random variable X by its values (or those of a concomitant variable) into K strata labeled $k = 1, 2, \ldots, K$, we can write the mean of X and the sample mean $\bar{X}$ as, respectively,

$$\mu = E(X) = \sum_{k=1}^{K} p_k E(X \mid X \in \text{strata } k); \tag{3.1}$$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \left(\sum_{\text{str } 1} X_i + \ldots + \sum_{\text{str } K} X_i \right) \tag{3.2}$$

$$= \sum_{\text{str } 1} \frac{n_1}{n} \frac{X_i}{n_1} + \ldots + \sum_{\text{str } K} \frac{n_K}{n} \frac{X_i}{n_K}, \tag{3.3}$$

where n_k = number of X_i's observed in strata, and p_k is the probability of being in strata.

Now, if the p_k's are known and we substitute them in (3.3) for n_k/n, we get $\bar{X}_{st}$, a stratified estimate. It will be biased, since the divisions of the sums in the populations are random; if the numbers observed in each population are controlled and taken to be np_1, np_2, etc., we have what is called a proportionally sampled estimate $\bar{X}_{ps}$ with

$$\text{var}(\bar{X}_{ps}) = \text{var}(\bar{X}) - \sum_{k=1}^{K} p_k (\mu_k - \mu)^2 / n, \tag{3.4}$$

where $\mu_k = E\{X \mid X \in \text{strata } k\}$. Thus, because of the use of prior analytic information we always get variance reduction.

The variance of $\bar{X}_{st}$ is not analytically tractable, but early studies reported in Kleijnen indicate it is close to (3.4) if n and all the n_k's are sufficiently large. Our simulation studies with stratifying α have confirmed this; because of the ease of computation of $P\{\alpha=1\}, P\{\alpha=2\}, P\{\alpha \geq 3\}$, as in section 2.3, it is natural to stratify α. Considerable variance reduction is obtained, especially for smaller values of ρ, the traffic intensity.

Unfortunately, when the quantity α is stratified in the bottom (denominator) of the estimator $W(n)$ very little overall variance reduction is obtained unless ρ is small. We do not understand this lack of variance reduction but apparently the correlation between the stratified version of $\bar{\alpha}(n)$ and $\bar{Y}(n)$ is reduced. Analytical studies of this effect are very difficult.

It is also possible to use a stratified estimate of $\bar{\alpha}$ in $W(n)$ and to control the top, $\bar{Y}(n)$, with say $C^{(6)}$ using $C^{(2)}$. This works well for small ρ, but _increases_ the variance as ρ approaches 1. Again it is difficult to understand this effect.

4. Conclusions and Summary

We have been able to obtain a worthwhile variance reduction using internal control variables, for the regenerative estimate of the limiting value of the mean waiting time in an M/M/1 queue. This reduction is obtained uniformly over all parameter values. It is fairly certain that the technique will work well with any GI/G/1 queue or other regenerative stochastic processes or systems. Internal stratified sampling schemes, however, did not work nearly as well.

The techniques can be extended to other stable stochastic systems, such as the Markov chains considered in Crane and Iglehart (1974b). In that case the computation of the mean values of the controls is simpler because of the structure of the Markov chain.

The main problem in applying the internal control variance reduction technique seems to lie in the fact that the estimator proposed by Crane and Iglehart (1974a) involves a ratio of two random variables, and these are difficult to work with in general.

An alternative which will be considered later is to use the existence of regeneration points more specifically to obtain variance reduction with the classical estimator $\widetilde{W}_m$ given at (2.1). One advantage which the regenerative estimator $W(n)$ has over $\widetilde{W}_m$ is the ease of obtaining confidence interval estimates or estimates of the precision of $W(n)$ and $\widetilde{W}(n)$. This is not a drawback if the simulation is large and more than one (say ten or twenty) realizations of the queue are obtained.

To fix ideas note that we can write $\widetilde{W}_m$ as

24

$$\widetilde{W}_m = \frac{1}{m} \sum_{j=0}^{m} W_j$$

$$= \frac{1}{m} \left\{ \sum_{k=1}^{N(m)} Y_k + Y'_{N(m)+1} \right\} ,$$

where $N(m)$ is the number of completed busy periods in the queue in $[0,m]$, Y_k as before is the sum of the waiting times in the $k\underline{th}$ cycle, and $Y'_{N(m)+1}$ the sum of the waiting times in the last, incomplete cycle.

Now it is possible to apply internal controls to each Y_k in the sum. Problems arise in estimating the coefficients β in the control because they involve a random sum of random variables. But it is much easier to find a control C for Y_k rather than the difference $Y_k - \alpha_k$, and also it is still possible to apply external controls as well as internal controls.

These ideas will be followed up in a later paper.

APPENDIX

To implement either the internal control or stratified sampling techniques certain theoretical parameters associated with the GI/G/1 queue are required. In this appendix we shall indicate the values of these parameters in so far as they can be calculated. These values are either well-known or easily calculated. For a reference to the known formulas see Cohen (1969).

We begin with $E\{\alpha_1\}$, the expected number of customers served in a busy period. For the general GI/G/1 queue recall that we let $X_n = v_{n-1} - u_n$ and $S_n = X_1 + \ldots + X_n$, for $n \geq 1$, with $S_0 = 0$. Then $\alpha_1 = \inf\{n > 0: S_n \leq 0\}$. The general expression for $E\{\alpha_i\}$ is given by

$$E\{\alpha_1\} = \exp\{\sum_{n=1}^{\infty} n^{-1} P[S_n > 0]\},$$

an impossible expression to evaluate in general. Another useful expression for $E\{\alpha_1\}$ is

$$E\{\alpha_1\} = 1/P\{W=0\}, \qquad (A.1)$$

where W is the stationary waiting time. In the special case of M/G/1, however, we have

$$E\{\alpha_1\} = (1+\rho)^{-1}.$$

Now for the queue GI/M/1 we can use (A.1) and the stationary distribution of the embedded Markov chain to conclude that

$$E\{\alpha_1\} = (1-\delta)^{-1},$$

26

where δ is the root inside the unit circle of

$$z - U\{\mu(1-z)\} = 0$$

with $U(s) = E\{e^{-su_1}\}$, Re $s \geq 0$, and u_1 is an exponential (μ) r.v. It is easy to check that $\delta = \rho$ for M/M/1 queues. Daley (1975) has recently proposed the approximation to δ given by

$$\gamma = a_1(1-\rho)^2 + 2(1-b^{-1})\rho + (2b^{-1}-1)\rho^2,$$

where $a_1 = P\{u_1=0\}$, $E\{u_1\} = 1$, and $b = E\{u_1^2\}$. This approximation gives good results in a number of examples calculated by Daley (1975) and may be useful for the purposes of this paper.

Next we turn to the computation of $P\{\alpha_1=1\}$ and $P\{\alpha_1=2\}$. For the GI/G/1 case we have

$$P\{\alpha_1=1\} = P\{S_1 \leq 0\}$$

and

$$P\{\alpha_1=2\} = P\{S_1>0, S_2 \leq 0\},$$

both of which can be worked out with a little effort. For the M/M/1 queue

$$P\{\alpha_1=1\} = (1+\rho)^{-1}$$

and

$$P\{\alpha_1=2\} = \rho(1+\rho)^{-3}.$$

For the M/G/1 queue

$$P\{\alpha_1=1\} = V(\lambda)$$

where $V(\lambda) = E(e^{-\lambda v_0})$, and for the GI/M/1 queue

27

$$P\{\alpha_1=1\} = 1 - U(\mu),$$

where $U(s)$ is given above. For the $M/E_k/1$ queue and the $E_k/M/1$ queue the value $P\{\alpha_1=2\}$ can be calculated with some effort. As these expressions are cumbersome they shall be omitted.

Next we turn to various partial expectations which are needed for internal control

$$E\{S_1^+ + S_2^+, \alpha_1 \geq 2\} = E\{S_1, S_1 > 0\} + E\{S_2^+, S_1 > 0\}$$

and

$$E\{\alpha_1, \alpha_1 \geq 2\} = E\{\alpha_1\} - P\{\alpha_1=1\}.$$

In the special case of the $M/M/1$ queue,

$$E\{S_1, S_1 > 0\} = \rho/\{\mu(1+\rho)\},$$

$$E\{S_2^+, S_1 > 0\} = [2\left(\frac{\rho}{1+\rho}\right)^2 + \frac{\rho^2}{(1+\rho)^3}]\,\mu^{-1},$$

and

$$E\{\alpha_1, \alpha_1 \geq 2\} = 2\rho/\{(1-\rho)(1+\rho)\}.$$

REFERENCES

[1]   COHEN, J. W. (1969). The Single Server Queue. North-Holland
      Publishing Co., Amsterdam.

[2]   CRAMÉR, H. (1946). Mathematical Methods of Statistics.
      Princeton University Press, Princeton.

[3]   CRANE, M. A. and IGLEHART, D. L. (1974a). Simulating stable
      stochastic systems, I: General multi-server queues. J. Assoc.
      Comput. Mach. 21, 103-113.

[4]   CRANE, M. A. and IGLEHART, D. L. (1974b). Simulating stable
      stochastic systems, II: Markov chains. J. Assoc. Comput.
      Mach. 21, 114-123.

[5]   CRANE, M. A. and IGLEHART, D. L. (1975a). Simulating stable
      stochastic systems, III: Regenerative processes and discrete-
      event simulations. Operations Res. 23, 33-45.

[6]   CRANE, M. A. and IGLEHART, D. L. (1975b). Simulating stable
      stochastic systems, IV: Approximation techniques. Management
      Sci. 21, 1215-1224.

[7]   DALEY, D. J. (1975). A simple approximation for the root of
      an equation in GI/M/s queues. Technical Report, Statistics
      Department (IAS), The Australian National University, Canberra.

[8]   GAVER, D. P. and THOMPSON, G. L. (1973). Probability and
      Programming Models in Operations Research. Brooks/Cole Pub.
      Co., Monterey, Calif.

[9]   IGLEHART, D. L. (1975). Simulating stable stochastic systems,
      V: Comparison of ratio estimators. Naval Res. Logist. Quart.
      22, 553-556.

[10] IGLEHART, D. L. (1976a). Simulating stable stochastic systems,
      VI: Quantile estimation. J. Assoc. Comput. Mach. 23, 347-360.

[11] IGLEHART, D. L. (1976b). Simulating stable stochastic systems,
      VII: Selecting the best system. To appear in Management Sci.

[12] KLEIJNEN, J. P. C. (1974). Statistical Techniques in Simula-
      tion, Part 1. Marcel Decker, New York.

[13] LEARMONTH, G. P. and LEWIS, P. A. W. (1973). Naval Postgraduate
      School random number generator package LLRANDOM. Naval Post-
      graduate School Report NPS55Lw73061A.

[14] LEARMONTH, G. P. and LEWIS, P. A. W. (1974). Statistical
      tests of some widely used and recently proposed uniform
      number generators. _Proc. Seventh Conference on Computer
      Science and Statistics_, Western Periodicals Co., North
      Hollywood, Calif.

[15] LEWIS, P. A. W., GOODMAN, A. S. and MILLER, J. M. (1969).
      A pseudo-random number generator for the system/360. _IBM
      Systems Journal_ 8, 136-146.

[16] MILLER, R. P. (1974). The jackknife--a review. _Biometrika_
      61, 1-15.

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
86-22		

4. TITLE (and Subtitle)	5. TYPE OF REPORT & PERIOD COVERED
Variance Reduction for Regenerative Simulations, I: Internal Control and Stratified Sampling for Queues	Technical Report
	6. PERFORMING ORG. REPORT NUMBER
Technical Report No. 86-22

7. AUTHOR(s)	8. CONTRACT OR GRANT NUMBER(s)
Donald L. Iglehart and Peter A. W. Lewis	N00014-72-C-0086

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Control Analysis Corporation 800 Welch Road, Palo Alto, CA 94304	NR-047-106

11. CONTROLLING OFFICE NAME AND ADDRESS	12. REPORT DATE
Operations Research Program Office of Naval Research Arlington, VA 20360	August 1976
	13. NUMBER OF PAGES

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report)
	Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

control variables, simulation, regenerative method, statistical analysis, stochastic systems, steady-state parameters, variance reduction.

are discussed

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We discuss two methods for reducing the variance of estimators of parameters of limiting distributions of stable stochastic processes in simulations. The methods are discussed in the context of the simple GI/G/1 queue. Of the two methods one, which we called an internal control variable, gives a variance reduction which is roughly uniform over values of the parameters of the process and, in particular, works well for values of ρ, the ——cont.

UNCLASSIFIED

cont.

traffic intensity, close to 1.